DOCUMENT RESUME

ED 459 813 IR 058 349

AUTHOR Rauber, Andreas; Muller-Kogler, Alexander
TITLE Integrating Automatic Genre Analysis into Digital Libraries.
PUB DATE 2001-06-00
NOTE 11p.; In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, VA, June 24-28, 2001). For entire proceedings, see IR 058 348. Figures may not reproduce well.
AVAILABLE FROM Association for Computing Machinery, 1515 Broadway, New York, NY 10036. Tel: 800-342-6626 (Toll Free); Tel: 212-626-0500; e-mail: acmhelp@acm.org. For full text: http://www1.acm.org/pubs/contents/proceedings/dl/379437/.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Content Analysis; Data Processing; *Electronic Libraries; *Information Retrieval; Information Systems; Library Services; Online Searching; Relevance (Information Retrieval); Visual Aids
IDENTIFIERS *Genre Approach

ABSTRACT

With the number and types of documents in digital library systems increasing, tools for automatically organizing and presenting the content have to be found. While many approaches focus on topic-based organization and structuring, hardly any system incorporates automatic structural analysis and representation. Yet, genre information (unconsciously) forms one of the most distinguishing features in conventional libraries and in information searches. This paper presents an approach to automatically analyze the structure of documents and to integrate this information into an automatically created content-based organization. In the resulting visualization, documents on similar topics, yet representing different genres, are depicted as books in differing colors. This representation supports users intuitively in locating relevant information presented in a relevant form. (Contains 22 references.) (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 459 813

Integrating Automatic Genre Analysis into Digital Libraries

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Andreas Rauber, Alexander Müller-Kögler
Department of Software Technology
Vienna University of Technology
Favoritenstr. 9 - 11 / 188
www.ifs.tuwien.ac.at/ifs

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

D. Cotton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

ABSTRACT

With the number and types of documents in digital library systems increasing, tools for automatically organizing and presenting the content have to be found. While many approaches focus on topic-based organization and structuring, hardly any system incorporates automatic structural analysis and representation. Yet, genre information (unconsciously) forms one of the most distinguishing features in conventional libraries and in information searches. In this paper we present an approach to automatically analyze the structure of documents and to integrate this information into an automatically created content-based organization. In the resulting visualization, documents on similar topics, yet representing different genres, are depicted as books in differing colors. This representation supports users intuitively in locating relevant information presented in a relevant form.

Keywords

Genre Analysis, Self-Organizing Map (SOM), SOMLib, Document Clustering, Visualization, Metaphor Graphics

1. INTRODUCTION

While the question of *what* a document is about has been recognized as being crucial for presenting relevant information to a user, the question of *how* a given piece of information is presented is largely neglected by most present electronic information systems. Yet, this type of information is – mostly unconsciously – used in almost any contact with information in everyday life. Personal letters are treated differently than mass-mailings, a short story is read on different occasions than long novels, popular science literature addresses a different readership than dissertations or scientific papers, both of which themselves will provide highly similar information at differing levels of detail for different audiences. Specific sub-genres, such as for example, executive summaries or technical reports, were even specifically designed to satisfy the same information need, i.e. to provide information about a given topic, in different ways. Whenever looking for information, these issues are taken into account, and they form one of the most important distinguishing features in conventional libraries, together with other non-content based information such as the age of a document, the fact whether it looks like it is being used frequently or remains untouched for long periods of time, and many others.

As long as a digital library can be cared for in a way similar to how conventional libraries are organized, this type of information is carefully captured in the form of metadata descriptions, and provided to the user, albeit mostly in rather inconvenient, not intuitive textual form. Yet, with this information available, ways for more intuitive representations can be devised. A different situation is encountered in many less-controlled digital library settings, where pieces of information from different sources are integrated, or the mere amount of information added to a repository effectively prevents it from being manually indexed and described. For these settings an automatic analysis of the structure of a given piece of information is essential to allow the user to quickly find the correct document, not only in terms of the content provided, but also with respect to the way this content is presented.

In this paper we present a way to provide automatic analysis of the structure of text documents. This analysis is based on a combination of various surface level features of texts, such as word statistics, punctuation information, the occurrences of special characters and keywords, as well as mark-up tags capturing image, equation, hyperlink and similar information. Based on these structural descriptions of documents, the self-organizing map (SOM) [12], a popular unsupervised neural network, is used to cluster documents according to their structural similarities. This information is incorporated into the SOMLib digital library system [17] which provides an automatic, topic-based organization of documents using again the self-organizing map to group documents according to their content. The libViewer, a metaphor-graphical interface to the SOMLib system depicts the documents in a digital library as hardcover or paperback books, binders, or papers, sorted by content into various bookshelves, labeled by automatically extracted content descriptors using the LabelSOM technique. Integrating the results of the structural analysis of documents allows us to color the documents, which are sorted by subject into the various shelves, according to their structural similarities, making e.g. complex descriptions stand apart

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
JCDL'01, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006 ...$5.00.

IR058349

BEST COPY AVAILABLE

from summaries or legal explanations on the same subject. Similarly, interviews on a given topic are depicted different from reports, as are numerical tables or result listings. We demonstrate the benefits of an automatic structural analysis of documents in combination with content-based classification using a collection of news articles from several Austrian daily, weekly and monthly news magazines.

The reminder of this paper is structured as follows. Section 2 provides a brief introduction into the principles of genre analysis and presents a review of related work in this area. We proceed by presenting the architecture and training procedure of the self-organizing map in Section 3. The application of the SOM to content-based document classification is presented in Section 4, including an overview of the various modules of the SOMLib digital library system with a special focus on the metaphor-graphics based libViewer interface. We then present our approach to structural classification of documents and its integration with the content-based representation provided by the SOMLib system in Section 5. Section 6 presents our experimental results using a collection of newspaper articles, reporting on content-based organization, structural classification, and their integration. Some conclusions as well as future work are listed in Section 7.

2. GENRE ANALYSIS

Genre analysis has a long history in linguistic literature. Conventionally, *genre* is associated with terms such as short stories, science fiction, novels of the 17th or 18th century, fiction, reports, satire, and many others. Still, the definition of genre is somewhat vague. According to Webster's Dictionary of English Language, genre is defined as *a category of artistic, musical, or literary composition characterized by a particular style, form, or content.* Although differing definitions may be found, the main goal of genre analysis is to identify certain subgroups within a set of given objects that share a common form of transmission, purpose, and discourse properties. Basically, the term *genre* can be applied to most forms of communications, although it is frequently restricted to non-interactive, and, for the scope of this paper, literary information, excluding music or film genres. While the common interpretation of genre refers to literary styles, such as *fiction, novel, letter, manuals*, etc., automatic analysis of genres takes a slightly different approach, focusing on structural analysis using surface level cues as the main structural similarity between documents, from which genre-style information is deducted..

Several approaches have been taken to evaluate the structure or readability of text documents, resulting in numerous different measures for grading texts automatically based on surface features. Many of these features are readily available in various implementations of the Unix *STYLE* command [6]. Among the measures included in this package are the *Kincaid Formula*, which is targeted towards technical material, having been developed for Navy training manuals. The *Flesh reading easy formula* stems from 1948 and is based on English school texts covering grades 3 to 12. A similar measure is the *SMOG-Grading*, or the *WSTF Index*, which has been developed specifically for German texts. All these measures basically compute their score by combining information about the number of words and syllables per sentence as well as characters per word statistics, weighted by various constants, to obtain the according grades.

More complex stylistic analyses can be found in the seminal work of Biber [1, 2]. He uses metrics such as pronoun counts and general text statistics to cluster texts in order to find underlying dimensions of variation and to detect general properties of genres.

More recently, classification of text documents by genre has been analyzed by Karlgren et al. [10]. Again, a number of different features are used to describe the structural characteristics of documents. However, additionally to the standard surface cues, additional features requiring syntactic parsing and tagged texts, such as required for noun counts, present participle count etc., were included. Discriminant analysis is used to obtain a set of discriminant functions based on a pre-categorized training set. This line of research is continued in [8], reporting in detail on the various features used for stylistic analysis. The stylistic variations of documents are further visualized as scatter plots based on combinations of two features. Specific areas are then (manually) assigned special genre-type descriptors to help users with analyzing the clusters of documents found in the scatter plot.

Recognizing the importance of integrating genre analysis into a content-based information retrieval process, the DropJaw interface [3, 9] incorporates genre-based classification using C4.5 based decision trees into content-based clustering using a hierarchical agglomerative group-average clustering algorithm. Documents are then represented in a two-dimensional matrix, with the rows representing the topical clusters found in the document set, whereas the columns organize these documents according to a number of genres the decision tree was trained to recognize.

A different approach describing documents by a number of facets rather than directly assigning a genre is reported in [11]. A facet is a property which distinguishes a class of texts that answers to certain practical interests, and which is associated with a characteristic set of computable structural or linguistic properties. Three principal categorical facets are analyzed. *Brow* characterizes a text with respect to the intellectual background required to understand a text, subdivided into *popular, middle, upper-middle*, and *high*. A binary *narrative* facet decides whether a text is written in a *narrative style*, and the third facet, *genre*, classifies a text either as *reportage, editorial, scitech, legal, nonfiction*, or *fiction*. A set of 55 lexical, character-level and derivative cues are used to describe the documents, and logistic regression is used to create a classifier based on a training set of 402 manually classified texts.

In [21], Ries applies genre classification to spontaneous spoken conversations, including features such as pauses in the conversation as well as histograms of a number of keywords, using a backpropagation-type neural network for the subsequent analysis.

It is interesting to note, that, although unsupervised methods are frequently used for content-based analysis of information, most of current research work turns to supervised models when it comes to the analysis of genre. This might be due to the case, that people tend to think in terms of well-defined genres, rather than in terms of structurally similar documents. Still, we find documents to frequently exhibit characteristics of several different genres to differing degrees. This is the more so as for hardly any genre there is a strict and well-defined, non-overlapping set of criteria by which it can be described, making strict classification as impossible a task as strict content-based classification. Similar as for

content-based document organization, unsupervised cluster analysis of genre-oriented document descriptions should be able to capture the structural similarities accordingly.

3. THE SELF-ORGANIZING MAP

The self-organizing map [13] provides cluster analysis by producing a mapping of high-dimensional input data $x, x \in \Re^n$, onto a usually 2-dimensional output space while preserving the topological relationships between the input data items as faithfully as possible. This model consists of a set of units, which are arranged in some topology where the most common choice is a two-dimensional grid. Each of the units i is assigned a weight vector m_i of the same dimension as the input data, $m_i \in \Re^n$, initialized with random values.

During each learning step, the unit c with the highest activity level, i.e. the *winner* c with respect to a randomly selected input pattern x, is adapted in a way that it will exhibit an even higher activity level at future presentations of that specific input pattern. Commonly, the activity level of a unit is based on the Euclidean distance between the input pattern and that unit's weight vector. The unit showing the lowest Euclidean distance between it's weight vector and the presented input vector is selected as the winner. Hence, the selection of the winner c may be written as given in Expression (1).

$$c : \|x - m_c\| = \min_i \{\|x - m_i\|\} \tag{1}$$

Adaptation takes place at each learning iteration and is performed as a gradual reduction of the difference between the respective components of the input vector and the weight vector. The amount of adaptation is guided by a learning rate α that is gradually decreasing in the course of time. This decreasing nature of adaptation strength ensures large adaptation steps in the beginning of the learning process where the weight vectors have to be tuned from their random initialization towards the actual requirements of the input space. The ever smaller adaptation steps towards the end of the learning process enable a fine-tuned input space representation.

As an extension to standard competitive learning, units in a time-varying and gradually decreasing neighborhood around the winner are adapted, too. Pragmatically speaking, during the learning steps of the self-organizing map a set of units around the winner is tuned towards the currently presented input pattern enabling a spatial arrangement of the input patterns such that alike inputs are mapped onto regions close to each other in the grid of output units. Thus, the training process of the self-organizing map results in a topological ordering of the input patterns.

The neighborhood of units around the winner may be described implicitly by means of a (Gaussian) neighborhood-kernel h_{ci} taking into account the distance—in terms of the output space—between unit i under consideration and unit c, the winner of the current learning iteration. This neighborhood-kernel assigns scalars in the range of $[0, 1]$ that are used to determine the amount of adaptation ensuring that nearby units are adapted more strongly than units farther away from the winner.

It is common practice that in the beginning of the learning process the neighborhood-kernel is selected large enough to cover a wide area of the output space. The spatial width of the neighborhood-kernel is reduced gradually during the learning process such that towards the end of the learning process just the winner itself is adapted. This strategy enables the formation of large clusters in the beginning and fine-grained input discrimination towards the end of the learning process.

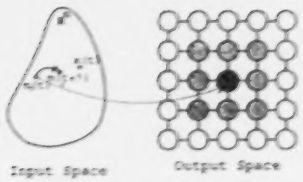

Figure 1: SOM architecture and training process

In combining these principles of self-organizing map training, we may write the learning rule as given in Expression (2). Please note that we make use of a discrete time notation with t denoting the current learning iteration. The other parts of this expression are α representing the time-varying learning rate, h_{ci} representing the time-varying neighborhood-kernel, x representing the currently presented input pattern, and m_i denoting the weight vector assigned to unit i.

$$m_i(t+1) = m_i(t) + \alpha(t) \cdot h_{ci}(t) \cdot [x(t) - m_i(t)] \tag{2}$$

A simple graphical representation of a self-organizing map's architecture and its learning process is provided in Figure 1. In this figure the output space consists of a square of 25 units, depicted as circles. One input vector $x(t)$ is randomly chosen and mapped onto the grid of output units. In the next step of the learning process, the winner c showing the highest activation is selected. Consider the winner being the unit depicted as the black unit in the figure. The weight vector of the winner, $m_c(t)$, is now moved towards the current input vector. This movement is symbolized in the input space in Figure 1. As a consequence of the adaptation, unit c will produce an even higher activation with respect to input pattern x at the next learning iteration, $t+1$, because the unit's weight vector, $m_c(t+1)$, is now nearer to the input pattern x in terms of the input space. Apart from the winner, adaptation is performed with neighboring units, too. Units that are subject to adaptation are depicted as shaded units in the figure. The shading of the various units corresponds to the amount of adaptation and thus, to the spatial width of the neighborhood-kernel. Generally, units in close vicinity of the winner are adapted more strongly and consequently, they are depicted with a darker shade in the figure.

4. THE SOMLIB SYSTEM

4.1 Feature extraction

In order to utilize the SOM for organizing documents by their topic a vector-based description of the content of the documents needs to be created. While manually or semi-automatically extracted content descriptors may be used,

research results have shown that a rather simple word frequency based description is sufficient to provide the necessary information in a very stable way [4, 14, 15, 19]. For this word frequency based representation a vector structure is created consisting of all words appearing in the document collection. This list of words is usually cleaned from so-called stop-words, i.e. words that do not contribute to content representation and topic discrimination between documents. Again, while manually crafted stop-word lists may be used, simple statistics allow the removal of most stop-words in a very convenient and language- or subject-independent way. On the one hand, words appearing in too many documents, e.g. in more than half of all documents, can be removed without the risk of loosing content information, as the content conveyed by these words is too general. On the other hand, words appearing in less than a minimum number of, say, 5 to 10 documents can be omitted for content-based classification, as the resulting sub-topic granularity would be too small to form a topic cluster of its own. Note, that the situation is different in the information retrieval domain, where rather specific terms need to be indexed to facilitate retrieval of a very specific subset of documents. In this respect, content-based organization and browsing of documents constitutes a conceptually different approach to accessing and interacting with document archives by browsing topical hierarchies. This obviously has to be supplemented by various searching facilities, including information retrieval capabilities as they are currently realized in many systems.

The documents are described within the resulting feature space of commonly between 5.000 and 15.000 dimensions, i.e. distinct terms, by the words they are made up of. While a basic binary indexing may be used to describe the content of a document by simply stating whether a word appears in the document or not, more sophisticated schemes, such as $tf \times idf$, i.e. term frequency times inverse document frequency [22], provide a better content representation. This weighting scheme assigns higher values to terms that appear frequently within a document, i.e. have a high term frequency, yet rarely within the complete collection, i.e. have a low document frequency. Usually, the document vectors are normalized to unit length to make up for length differences of the various documents.

4.2 Topic-based organization

The resulting vector representations are fed into a standard self-organizing map for cluster analysis. As a result, documents on similar topics are located on neighboring units in the two-dimensional map display. In the simplest form, a document collection may then be represented as a rectangular table with similar documents being mapped onto the same cells. Using this model, users find a document collection to be automatically structured by content in a way similar to how documents are organized into shelves in conventional libraries. Due to its capabilities of automatically structuring a document collection by subject, we have chosen the SOM as the basic building block of our SOMLib digital library system [19]. Enhanced models of the SOM, such as the growing hierarchical self-organizing map (GHSOM), further allow the automatic detection of topical hierarchies by creating a layered structure of independent SOMs that adapt their size accordingly [20].

4.3 Labeling

While the SOM found wide appreciation in the field of text classification, its application had been limited by the fact that the topics of the various cluster were not evident from the resulting mapping. In order to find out which topics are covered in certain areas of the map, the actual articles had to be read to find descriptive keywords for a cluster. To counter this problem, we developed the LabelSOM method, which analyses the trained SOM to automatically extract a set of attributes, i.e. keywords, that are most descriptive for a unit [16]. Basically, the attributes showing a low quantization error value and a high weight vector value, comparable to a low variance and a high mean among all input vectors mapped onto a specific unit, are selected as labels. Thus, the various units are characterized by keywords describing the topics of the documents mapped onto them.

4.4 Visualization

Last, but not least, while the spatial organization of documents on the 2-dimensional map in combination with the automatically extracted concept labels supports orientation in and understanding of an unknown document repository, much information on the documents cannot be told from the resulting representation. Information like the size of the underlying document, its type, the date it was created, when it was accessed for the last time and how often it has been accessed at all, its language etc. is not provided in an intuitively interpretable way. Rather, users are required to read and abstract from textual descriptions, inferring the amount or recent-ness of information provided by a given document by comparing size and date information.

We thus developed the libViewer, a metaphor-graphics based interface to a digital library [18]. Documents are no longer represented as textual listings, but as graphical objects of different *representation types* such as binders, papers, hardcover books, paperbacks etc, with further metadata information being conveyed by additional metaphors such as *spine width*, *logos*, *well-thumbed spines*, different degrees of *dustiness*, *highlighting glares*, *position in the shelf* and others. Based on these metaphors we can define a set of mappings of metadata attributes to be visualized, allowing the easy understanding of documents, similar to the usage of Chernoff faces for multidimensional space representation [5]. Figure 2 provides an example of the visualization of documents in a digital library using the libViewer interface. Documents are depicted using different document type representations, with additional metadata being conveyed by their position, color, spine width, the logos and textual information depicted on the spine, dust or highlighting glares, well-thumbed bindings and others. Metadata information about documents thus can be easily interpreted and compared across a collection, with a larger amount of information being represented as compared to standard textual descriptions.

5. STRUCTURE AND GENRE ANALYSIS

5.1 Structural features

Although the content-based organization and metaphor-graphical visualization of documents provided by the SOMLib system greatly supports the user in interacting with a digital library, all meta-information about documents has to be created and provided manually. While the size of

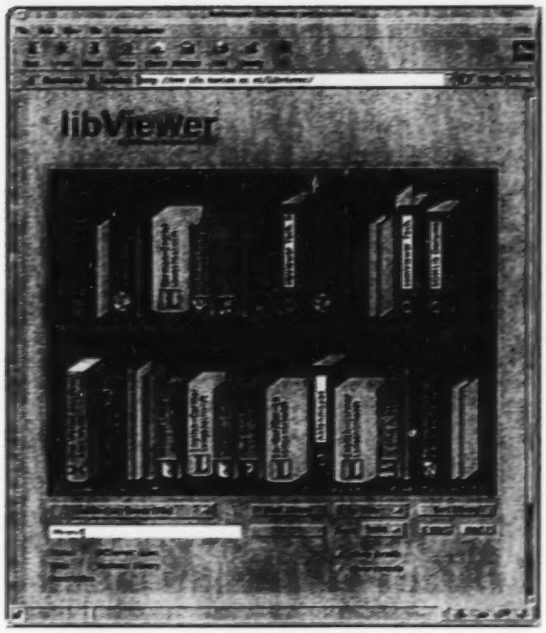

Figure 2: libViewer visualization of a digital library

a document, its date of creation, the date of last access, or the author are usually available, hardly any consistent genre-information is provided in most electronic document collections that have too high a volume of documents added frequently to allow manual classification. Yet, this type of information is important for a user to be able to pick the most appropriate document.

Similar to the task of content-based organization, we would like to have a way to automatically organize and categorize documents by their structure and stylistic features into basic types of genre, and to fuse this information with the content-based organization provided by the SOMLib system and its libViewer interface. Yet, we do not want to use supervised models, both because of the limitations introduced by the supervised process and because of the tedious task of having to produce an accurate training set manually. We thus propose to follow an approach similar to the one chosen for content-based document organization creating a feature vector representation of documents capturing the stylistic characteristics of documents. Clustering documents according to their similarity conveyed by their structural features should then reveal basic types of documents, or genres.

5.1.1 *Surface level cues*

For the set of features we have to restrict ourselves to those features that can be automatically extracted from documents with acceptable computational costs. Furthermore, similar to content-based classification, we want the process to be as language- and domain independent as possible to allow flexible application of the system in different settings. Basically, four distinct types of features can be distinguished, which are (1) text complexity information and text statistics, (2) special character and punctuation counts, (3) characteristic keywords, and (4) format-specific markups.

5.1.2 *Text complexity measures*

Text complexity measures are based on word statistics such as the average number of words per sentence, the average word length, number of sentences and paragraphs, and similar formatting information. While being rather simple metrics, the complexity of certain constructs turns out to be captured quite well by these measures, especially if combined with other characteristics such as punctuation marks. More extensive measures analyzing the nesting depth of sentences may be considered, although we did not integrate them into the experiments presented below. Furthermore, instead of using the basic measures, derived measures like any of the existing readability grade methods, such as Kincaid, Coleman-Liau, Wheeler-Smith Index, Flesh Index, may be used as more condensed representations. Still, these formulas are all based on the above-mentioned basic measures anyway, using various transformations to obtain graded representations according to stylistic evaluation parameters. We thus refrained from using those in our initial experiments, although they may be considered for follow-up evaluations.

5.1.3 *Special character and punctuation counts*

A wealth of stylistic information can be obtained from specific character counts. As most prominent among these we consider punctuation marks, as some of them are rather characteristic for certain text genres. For example, the presence of exclamation mark (*!*) is highly indicative of more emotional texts as opposed to pure fact reports in a news magazine setting. Interviews exhibit a rather high count of question marks and colons, whereas high counts for semicolons or commas are indicative of more complex sentence structures, especially if co-occurring with rather high sentence lengths. Otherwise, if co-occurring with rather short sentences they might rather be attributed to listings and enumerations of information items. Similarly, we have to analyze the occurrence of quotation marks, hyphens, periods, apostrophe, slash marks, various brackets and others. More complex punctuation information might be worth considering if it can be feasibly extracted from the given texts, such as ellipsis points, differences between single and double quotation marks, or regarding the usage of dashes versus hyphens, especially if the semantic context can be deducted. However, as these more complex features could not be extracted from the given text base integrating different sources we had to refrain from representing this level of structural information for the given experiments.

Other characters worth incorporating into stylistic analysis are financial symbols like *$*, *£*, *euro*. Furthermore, numbers as well as mathematical symbols, copyright and paragraph signs are worth including as they do hint at special categories of information, be it, for example, technical discussions, price listings, legal information, or simply examples to clarify and expand on a given topic.

5.1.4 *Stopwords and keywords*

Contrary to the principles of content representation, a lot of genre-information is conveyed by stop-words such as pronouns or adverbs. Thus, a list of characteristic words is added to the list of features, containing words such as *I*, *you*,

BEST COPY AVAILABLE

we, me, us, mine, yours or *much, little, large, very, highly, probably, mostly, certainly, which, that, where* and others. Please note that this list is language dependent, yet may be easily adapted for most languages. (These adaptions may be rather complex for specific languages using word inflections rather than specific keywords for some characteristic sentence structures.) Depending on the specific document collection to be considered and the desired focus of genre analysis, additional keywords may be added to that list to facilitate, for example, the recognition of fact-reporting versus opinion-modifying articles, or the separation of speculative articles.

5.1.5 Mark-up tags

The fourth group of features is formed by specific mark-up tags that are used to extract information about the document. Using such mark-up tags, information such as the amount of images present in a given document, the number of tables and equations, links, references, etc. can be extracted and included in the genre analysis. These obviously have to be adapted to the actual formatting of the source documents in such a way that they are consistently mapped, such as, e.g. mapping the *\begin{figure}* and the *IMGSRC* tags onto the image count feature for LaTeX and HTML documents.

Overall, the parser currently recognizes almost 200 different features, some of which are specifically geared towards a specific file format, such as HTML documents, whereas others are generally applicable. Depending on the goal of the structure analysis, only a subset of the available attributes may be selected for further analysis.

The resulting document descriptions are further fed into a self-organizing map for training. As a result of the training process documents with similar stylistic characteristics are located close to each other on neighboring units of the SOM. We may thus find longer documents with rather complex sentences in one area of the map, whereas in another area interviews, characterized by shorter sentences with a high count of colons and quotation marks might be located. While this map may now serve as a kind of genre-based access to the document archive, it needs to be integrated into the content-based library representation to support users in their information finding process.

5.2 Integrating content and structure based classification

Two possibilities offer themselves for the integrated representation of content- and structure-based organization within the SOMLib system. We can either chose a semi-automatic approach by assigning a specific document representation type to a certain region on the genre-map, such as, for example, assigning all documents in the upper right area of this map the representation type *binder*, whereas all documents in the area on the lower left part of the map may be depicted as *hardcover* documents. This approach allows a very intuitive representation of documents in the content-based representation if a sensible assignment of the genres identified by the map to the available representation types can be defined. However, this approach has shown to have several deficiencies when applied to large and unknown document collections. Firstly, the number of available representation types is rather limited as opposed to the number of different document types present in any collection. While the total number of 4 representation types available in the libViewer system so far may be supplemented with additional object types showing, for example, additional bindings, the total possible number still will be rather limited. Furthermore, as the available representation types are strictly distinct from each other, no gradual shift from one type of document to the other can be conveyed, thus actually forcing the documents to be definitely of one genre or the other. Providing more subtle information about its general structure, which might be well in between two specific genres, would be more appropriate and highly preferable.

Secondly, a rather high manual effort is required to analyze the actual genres identified by the SOM to provide a sensible mapping of genre areas on the map onto the respective graphical metaphors. Yet, the precise mapping is crucial for the usability of the classification result as users will associate a specific type of information with a certain representation metaphor. The assignment of a wrong representation template may thus turn out to be highly counterproductive for information location.

We thus favour an automatic approach for integrating the information provided by the genre map into the content-based visualization using the color-metaphor. Documents of similar structure shall be assigned a similar color to allow intuitive recognition and interpretation of structural similarities. This metaphor turns out to be almost perfectly suited for conveying the desired information as it does not transport any specific meaning in the given setting by itself, as opposed to the realistic document representation types, which are intuitively associated with a certain kind of information.

A rather straight-forward mapping of the position on the genre-map onto a specific color is realized by mapping the rectangular map area onto a plane of the RGB color cube, similar to the color-coding technique for cluster identification in SOMs [7]. Thus, documents mapped onto neighboring units on the genre map will be depicted in similar colors, allowing easy recognition of mutal similarity in style as well as depicting even gradual transitions between the various structural clusters. On the other hand, documents in different regions on the genre map, exhibiting a clearly distinct structure, are thus depicted in different colors on the content-based libViewer visualization.

6. EXPERIMENTS

6.1 Data set

Various series of experiments have been performed in different settings, including technical documents and web site analysis. For the experiments presented below we created a collection of news reports by downloading the web-editions of 14 daily, weekly, or monthly Austrian newspapers and print magazines. This setting exhibits several characteristica typical for digital libraries that cannot be tendered to manually as carefully as necessary. Information from different sources having different internal classification schemata is integrated. As the majority of documents stems from daily newspapers, the number of articles to be organized is too large to allow manual classification. Furthermore, while the topics covered by the various sources overlap, the perspectives from which these issues are presented differ. To a large degeree this can be attributed to the general genre of a source, such as newspapers and magazines specializing

on economic issues, but also, to a large degree, to the style of report chosen. In many situations we will find the same topic to be covered by a news report as well as by an interview or a column, or we find the same issue covered both in the general news as well as in, say, the economic section of a paper.

A cleansing procedure was implemented for each data source to automatically remove characteristic formatting structures of the various sources such as banners, footers, or navigation bars, as these would unduely interfere with the stylistic analysis. Furthermore, different HTML encodings for special characters were converted to a uniform representation. The results reported below are based on a subset of the entire collection consisting of 1.000 articles from March 2000. To keep the system as flexible and generally applicable as possible, no language- or domain-specific optimizations, such as stemming or the use of specific stop-word lists, were performed. The articles were parsed to create both a content-based and a structure-based description of the documents, which were further fed into two separate self-organizing maps for cluster analysis and representation. Due to space restrictions we cannot provide detailed representations of the according maps. Rather, we have selected representative clusters for detailed discussion.

6.2 Content-based organization

A 5×10 SOM was used for topical organization of the articles based on a 1.975-dimensional feature vector representation. The main topical clusters identified in the collection are, on the one hand, economic articles, which consist of several subclusters, such as a rather dominant group of articles relating to the telecom business, or the privatization of Austria's state-owned enterprises. This cluster is located in the lower left corner of the resulting SOM. On the opposite, upper right corner we find mostly articles covering political issues, such as the discussions concerning the formation of the new government following the 1999 elections. This political cluster basically covers the whole left area of the content-based SOM, moving from the initial elections-based discussions to the various political topics. Another prominent cluster is formed by sports reports covering soccer, formula 1, and horse races, to name a few. Other, smaller clusters address different areas of science, with two of the more prominent sub-clusters among these being devoted to medicine, and internet technologies.

Using the LabelSOM method appropriate labels were automatically extracted, describing the various topical clusters. (The keywords have been translated into English for discussion in the following sections.) We find, for example, one of the clusters representing articles on Austria's Freedom Party to be labelled with *fp, joerg, haider, haiders*, listing the parties abbreviation as well as the name of its political leader. The labels for this unit also demonstrate one of the weaknesses of the crude indexing approach chosen. As we do not apply any language-specific stemming techniques, the trailing genitiv-*s* causes the term *haider* to appear in two forms. Yet, this impreciseness does not cause distortions to the resulting content representation and organization, although language-specific adaptions would further improve the resulting classification, albeit sacrificing the language and domain independence.

This unit is located next to another unit labelled *minister of defence, fpoe, fp, westenthaler, klestil* in the bottom right corner, listing again the freedom party, another one of its leading polititians, Peter Westenthaler, as well as the name of Austria's president, Thomas Klestil. This co-location of similar, yet not identical topics, is one of the most important characteristics of SOMs making them particularly suitable for the organization of document collections for interactive browsing.

Shifting to another topic we find units from the economic cluster in the lower right corner to be labelled with, for example, *austria, stock-exchange, fonds manager, telecom* for the previously mentioned telecom-cluster, or *enterprise, state, va-tech, steel, oeiag, grasser, leitl* for the cluster on the privatization of Austria's state-own steel enterprise VA-Tech, and two of the leading polititians involved in the privatization process, Karl-Heinz Grasser and Christoph Leitl. Above this unit we find a similar topic, namely the privatization of the Austrian postal services, labelled with *privatization, psk, contracts*, nicely showing the topological ordering in the map.

As an example for labels from the sports cluster we might mention the cluster on Formula 1 with labels *races, bmw, williams, jaguar, wm*. Since we do not specify a manually designed stop-word list, some stop-words remain in the list of index terms and actually show up as labels as they form a prominent common feature of articles in a cluster. We thus also find labels such as *friday* and *both* as labels for the sports cluster. Again, a hand-crafted or semi-automatic approach may provide a better removal for stop-words, yet sacrificing domain and language independence to some degree, whereas the current approach can be applied to any given document collection. The cluster representing documents on soccer is labelled with *goal, real, madrid, muenchen, bavaria, rome, group*, listing important soccer clubs playing against each other in a given group of the tournament. Neighboring this unit we find another unit on soccer, this time labelled *champions league, cup, barcelona, madrid, porto*.

This map serves as a content-based index to the digital library, allowing users to find, by reading the labels, which topics are covered in which section of the library. The documents can be represented as located in an HTML table with the labels provided as text in the table cells. They may also be transformed into a graphical libViewer representation, with the according source of the document, i.e. the magazine's title etc. being provided as a logo on the spine. Yet, we do not have any information from the resulting representation whether a given document represents, say, an interview, a result listing etc., as this information is not provided as metatags within the articles.

6.3 Incorporating genre information

6.3.1 A genre map

While the content-based SOM provides an organization of articles by their subject, the genre SOM analyzes the structural features of the documents and groups the documents accordingly.

We find, for example, a rather dominant cluster representing various forms of interviews, moving from reports with several quotations in them to long interviews on a given subject. The labels extracted by the LabelSOM technique help us to identify the most distinctive characteristics of a given cluster. For the interviews we find the characteristic attributes to be the number of opening and closing quotes as

well as the colon. Further distinctions can be made by the average length of sentences as well as frequent line-breaks, setting interviews apart from articles with longer citations. Another cluster of documents having a high colon count, yet a completely different sentence length structure plus several other special characters occurring frequently in the text, such as opening and closing breaks or slashes, is formed by sports articles providing only result listings. These documents obviously also exhibit an unproportionally high count of numbers. While short reports separate themselves from longer articles due to a number of text length parameters, we find a further distinction within the short reports cluster having a higher count of numbers, yet less than for sports results. These articles are mostly reports or announcements for radio or TV shows, which may be either special documentaries or sports transmissions. Another large cluster of documents consists of legal documents, which set themselves apart by the frequent usage of the paragraph character (§). Internet articles are characterized by the *at-sign* (@).

6.3.2 *Integrating genre information into the topical organization*

While the labels extracted by the LabelSOM technique help the user interpreting and understanding what the SOM has learned, they are not sufficient to allow the user to intuitively tell.which cluster of documents corresponds to which specific genre. This is because the extracted low-level features do not correspond to what the casual user will attribute to characteristic for, say, an interview or sports results listings. Instead of assigning every unit to a specific genre, we map the genre SOM into an RGB-color space, such that documents located in the upper left corner are colored black, whereas the upper right corner is assigned green, the lower left corner red, and the lower right corner yellow. The units inbetween are automatically assigned intermediate colors. According to its structure, each document is now assigned a color based on its location on the genre SOM. This color is used for representing the document in the content-based libViewer representation. We thus may now expect to find documents on the same topic, which are located on the same shelf on the content-based SOM, to be coloured differently if they exhibit a different structure.

Figure 3 shows one shelf from the upper midle area of the SOM representing articles on soccer from the sports section. The general topic of this unit is given as shelf labels. As can be told from the logos, the documents on this unit stem mostly from the daily newspaper *Die Presse*, with only one article being from *Kurier*, another daily newspaper. Also, the average length of all articles on this unit is rather homogenous, with all documents being short articles, thus depicted with only the minimum spine width. Still, we can see from the differing coloring that several distinct types of documents are located on this unit. The first and the last article on this shelf, colored orange, both represent short result reports, listing only the outcomes of the various matches. The second document, colored in bright yellow, contains a rather emotional report, looking more like a transcript from a live broadcast, but not reporting explicitly on results, next to a green document providing a rather factual report on the same match in a rather complicated style. The one-but-last document, colored dark-red, contains a somewhat longer report on several matches, listing not only results, but also short descriptions of various sections of the matches. More

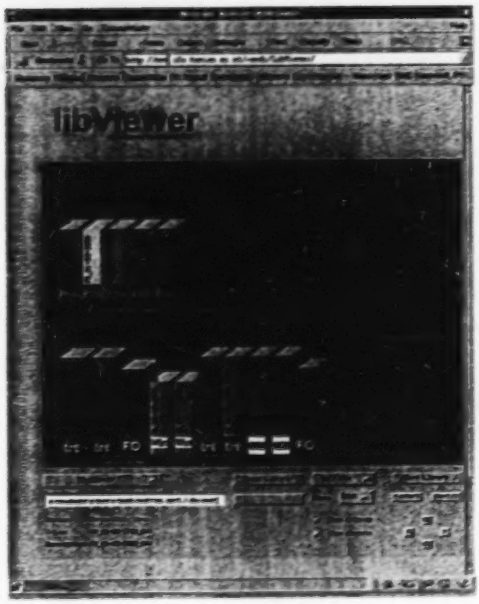

Figure 3: libViewer: sports and economy sections

important, however, is the fact that it also contains a report on the financial situation of one of the soccer clubs, thus being colored entirely different than the other soccer result reports. This difference in structure, and its partial membership in the economic articles genre, can be attributed to the frequent occurrence of the Euro currency symbol.

In the shelf beneath the soccer reports we find documents reporting on financial issues, or more precisely, on interest rates, representing articles from 4 different publications, one of which is a daily newspaper, 1 weekly, and 2 monthly magazines. Except for the weekly magazine *Format*, which is a general news magazine, all publications on this shelf have a strong focus on economic issues. The according labels are given as *banks*, *interest rates*. Two distinct types of articles can be distinguished on this unit, colored dark brown (the first three documents) and various shadings of green, respectively. When taking a look at the according documents, we find the dark brown ones to be rather complicated, extensive reports on interest rates issues, whereas the green documents are written in a rather informal style. These are mainly made up of short sentences and rethoric questions, and list the most important issues in a tabular form rather than by complex explanations. Please note, that the more complex articles also are longer, as can be seen from the wider spine width.

Figure 4 depicts the next two shelves down the row, continuing in the economic section of the map. Here we find a number of articles on the stock exchange in the upper shelf, whereas the lower shelf contains reports on the economic data from the print magazine *Die Wirtschaftswoche*. The first two documents on the lower shelf are colored in black, and they provide detailed percentual listings of the magazines subscriber structured, their average income etc. The third, green document reports on the same issue, yet rather represents a short overview article describing the general

BEST COPY AVAILABLE

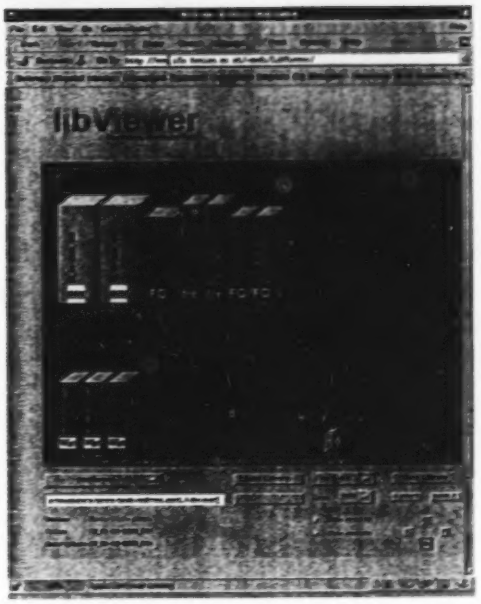

Figure 4: libViewer: economy-section (cont.)

results from the market study in simple terms. It thus is very similar in structure to the green documents discussed previously.

We also find a number of green documents in the upper shelf, providing short descriptions and buying recomendations for fonds and insurance policies based on fonds. The third, dark-brown document again describes a series of issues related to stock market transaction in a more complex structure, as can be expected from documents exhibiting this color. The first two documents, apart from being rather long, and thus depicted with rather broad spines, differ from the other reports by describing several companies and their performance by citing experts. While being rather detailed, we find many short sentences, enclosed by quotes, as the characteristic features of these documents, thus separating them from the neighboring darker lengthy description.

6.4 Evaluation

Unfortunately, the actual genre of a document cannot be intuitively told by the color it is represented in, nor could we find a way of how the different structural characteristics of documents could be automatically translated into a small set of distinct genres, which could then be represented by more intuitive metaphors (such as papers for interviews, harcover books for lengthy reports and paperbacks for shorter, simpler depictions). Although such a mapping would be possible in principle in a semi-automatic way by assigning different representation metaphors to different areas of the genre SOM, we prefer the automatic mapping of the structural position of a document on the genre SOM into a simple color space. While this metapher needs to be learned to be interpreted correctly, the actually effort required to understand the structure intuitively, rather than explicitly, has shown to be rather small and straight-forward. Furthermore, the chosen appröach allows gradual changes between various genres.

No large-scale usability study has yet been performed, although first tests with a small set of users, mostly students, have turned out encouraging results. After visiting a few documents on the respective areas of interest, most people had a feeling of what to expect from a document in a specific color, although they obviously were not able to describe it in terms of the low-level features used for classification by the map. Still, users know what to expect from, say, yellow to ochre documents (interviews, from black (numerical listings), greenish (short, simple articles), or others.

Obviously, the proposed approach to structural analysis will not be able to provide a full-scale genre analysis, capturing the fine differences between certain types of information representation, especially if they involve high-level linguistic analysis. To provide a rather far-fetched example, the presented system will definitely fail to separate a satire from factual information. It thus does not perform genre analysis in the strict sense.

Yet by capturing structural characteristics and similarities clearly is able to uncover specific genre information in given settings. While this might lead to misunderstandings in some situations, such as the impossibility of telling factual from fictional information (genre-wise), it should provide considerable support to the user trying to satisfy an information need. This is especially true as the utilization of the self-organizing map to produce a topology-preserving mapping allows to capture gradual differences between various structural concepts in a straight-forward manner.

7. CONCLUSIONS AND FUTURE WORK

Providing structural information about documents is essential to help users decide about the relevance of documents available in a digital library. Most document collections thus try to convey this information by using carefully designed metadata describing the genre of a resource. However, in many cases this uniform description of documents cannot be provided manually. This is especially true for digital libraries integrating documents from different sources, or where the number of documents to be described effectively prohibits manual classification.

In this paper we presented an automated approach to the structural analysis of text documents. Characteristic features such as the average length of a sentence, counts of punctuation marks and other special characters, as well as specific words such as pronouns etc. can be used to describe the structural characteristics of a document. The self-organizing map, a popular unsupervised neural network, is used to cluster the documents according to their similarity. Documents are then colored according to their location on the resulting two-dimensional map, such that structurally similar documents are colored similarly.

The result of the structural analysis is further incorporated into the content-based organization and representation of a digital library provided by the SOMLib system. Documents on the same topic, yet providing a different perspective of the same subject, such as reports and interviews, complex analyses, or short descriptions, are thus shown as books of different colors in the resulting graphical representation provided by the libViewer interface.

Initial experiments have shown encouraging results. In the next steps we would like to refine the mapping from the position on the structural SOM onto an appropriate color in such a way that the mutual similarity or distance of two units is reflected in the perceived distance of the colors assigned to the according documents. This would allow the structural similarities between, say, different types of interviews, to be more evident by assigning somewhat more similar colors to documents that are part of a larger cluster consisting of several sub-clusters. Such an improved mapping can be achieved by using distance information provided by the weight vectors of the units in the self-organizing map. Furthermore, a more suitable color space in terms of human perception may be chosen to further increase the perceived similarities and dissimilarities.

Secondly, we will take a closer look at additional features that offer themselves for genre analysis. First experiments indicate that, for example, a distinction between fact-reporting articles versus opinion-forming articles is possible by including additional keywords in the list of features.

8. REFERENCES

[1] D. Biber. *Variations across Speech and Writing*. Cambridge University Press, UK, 1988.

[2] D. Biber. A typology of english texts. *Linguistics*, 27:3 – 43, 1989.

[3] I. Bretan, J. Dewe, A. Hallberg, N. Wolkert, and J. Karlgren. Web-specific genre visualization. In *Proc of WebNet '98*, Orlando, FL, November 1998. `http://www.stacken.kth.se/~dewe/`.

[4] H. Chen, C. Schuffels, and R. Orwig. Internet categorization and search: A self-organizing approach. *Journal of Visual Communication and Image Representation*, 7(1):88–102, 1996. `http://ai.BPA.arizona.edu/papers/`.

[5] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal American Statistical Association*, 68:361–368, 1973.

[6] L. Cherra and W. Vesterman. Writing tools: The STYLE and DICTION programs. Technical Report 91, Bell Laboratories, Murray Hill, NJ, 1981. Republished as part 4.4BSD User's Supplementary Documents by O'Reilly.

[7] J. Himberg. A SOM based cluster visualization and its application for false coloring. In *Proc Int'l Joint Conf on Neural Networks (IJCNN 2000)*, Como, Italy, July 24. – 27. 2000. IEEE Computer Society.

[8] J. Karlgren. Stylistic experiments in information retrieval. In T. Strzalkowski, editor, *Natural Language Information Retrieval*. Kluwer, 1999. `http://www.sics.se/~jussi/Artiklar/`.

[9] J. Karlgren, I. Bretan, J. Dewe, A. Hallberg, and N. Wolkert. Iterative information retrieval using fast clustering and usage-specific genres. In *Proc Eighth DELOS Workshop on User Interfaces in Digital Libraries*, pages 85–92, Stockholm, Sweden, October 1998. `http://www.stacken.kth.se/~dewe/`.

[10] J. Karlgren and D. Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proc 15. Int'l Conf on Computational Linguistics (COLING '94)*, Kyoto, Japan, 1994. `http://www.sics.se/~jussi/Artiklar/`.

[11] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In *Proc 8. Conf Europ. Chapter of the Association for Computational Linguistics (ACL/EACL97)*, pages 32–38, Madrid, Spain, 1997. `http://spell.psychology.wayne.edu/~bkessler/`.

[12] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 1982.

[13] T. Kohonen. *Self-organizing maps*. Springer-Verlag, Berlin, 1995.

[14] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, and A. Saarela. Self-organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, May 2000. `http://ieeexplore.ieee.org/`.

[15] D. Merkl and A. Rauber. Document classification with unsupervised neural networks. In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval*, pages 102–121. Physica Verlag, 2000. `http://www.ifs.tuwien.ac.at/~andi/LoP.html`.

[16] A. Rauber. LabelSOM: On the labeling of self-organizing maps. In *Proc Int'l Joint Conf on Neural Networks (IJCNN'99)*, Washington, DC, July 10. - 16. 1999. `http://www.ifs.tuwien.ac.at/~andi/LoP.html`.

[17] A. Rauber. SOMLib: A digital library system based on neural networks. In E. Fox and N. Rowe, editors, *Proc ACM Conf on Digital Libraries (ACMDL'99)*, pages 240–241, Berkeley, CA, August 11. - 14. 1999. ACM. `http://www.acm.org/dl`.

[18] A. Rauber and H. Bina. Visualizing electronic document repositories: Drawing books and papers in a digital library. In *Advances in Visual Database Systems: Proc IFIP TC2 Working Conf on Visual Database Systems*, pages 95 – 114, Fukuoka, Japan, May. 10.- 12. 2000. Kluwer Academic Publishers. `http://www.ifs.tuwien.ac.at/~andi/LoP.html`.

[19] A. Rauber and D. Merkl. The SOMLib Digital Library System. In*Proc 3. Europ. Conf on Research and Advanced Technology for Digital Libraries (ECDL99)*, LNCS 1696, pages 323–342, Paris, France, September 22. - 24. 1999. Springer. `http://www.ifs.tuwien.ac.at/~andi/LoP.html`.

[20] A. Rauber, M. Dittenbach, and D. Merkl. Automatically detecting and organizing documents into topic hierarchies: A neural-network based approach to bookshelf creation and arrangement. In *Proc 4. Europ. Conf on Research and Advanced Technologies for Digital Libraries (ECDL2000)*, LNCS 1923, pages 348–351, Lisboa, Portugal, September 18. - 20. 2000. Springer. `http://www.ifs.tuwien.ac.at/~andi/LoP.html`.

[21] K. Ries. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proc Europ. Conf on Speech Communication and Technology (EUROSPEECH99)*, Budapest, Hungary, September 5-9 1999.

[22] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

BEST COPY AVAILABLE